JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

BIG DATA AND PREDICTIVE ANALYTICS IN MANUFACTURING ENTERPRISES FOR ENHANCED DECISION MAKING

[1]Rohit Katkar, [2]Dr. Rajesh Buktar
[1]M. Tech. student, [2]Professor
[1]Mechanical Engineering,
[1]Sardar Patel College of Engineering, Mumbai, India

Abstract: In the fourth industrial revolution, smart manufacturing is reshaping the manufacturing in the industries. The use of IoT technologies facilitates machine-to-machine communication and the flow of information from source to system, which results in a vast amount of machine data getting generated. Using this data enterprises can gain actionable insights & make better data-driven decisions. Big Data Analytics is one of the emerging & advanced analytic industry4.0 technologies, which help to process large & diverse data sets. Today's competitive environment forces enterprises to process this high-speed data & drive new opportunities. This paper starts with the definition and characteristics of big data, its sources, and its format. Further, it gives a brief explanation of the Hadoop ecosystem which helps us to understand a suite of services available on the cloud to solve big data problems. Big data technologies provide an opportunity to deploy predictive maintenance. Hence, in this paper, we have leveraged big data analytics for performing predictive analytics using a public dataset. Graphs have been plotted for various variables which help to bring visibility in the frequency of failure in advance & hence save the unplanned downtime. These graphs help to understand the range of any variable from which their machine component fails. In the end, an overview of the benefits of big data analytics in manufacturing has also been presented.

Index Terms - Big data, big data analytics, Hadoop ecosystem, Spark, Machine learning, predictive maintenance.

I. INTRODUCTION:

The fourth industrial revolution or industry 4.0 is the ongoing automation of traditional manufacturing and industrial practices, using modern smart technology. Large-scale machine-to-machine communication (M2M) and the Internet of Things (IoT) are integrated for increased automation, improved communication, and self-monitoring, and production of smart machines that can analyze and take corrective action without the need for human intervention (1). In industry 4.0 or simply I4, we connect the physical system with the cyber system called a cyber-physical system (CPS) (2). As the fourth industrial revolution makes use of the IoT system to connect both systems, it produces data. Data are a set of values of qualitative or quantitative variables from one or more objects or machines. Data are collected, measured, reported, and analyzed using graphs, images, charts, and other modes of analysis (3). When this data comes in large quantities, we call it 'Big Data.

The major contribution of the paper is:

1. Overview of the Hadoop ecosystem
2. Big data analytics applications, its benefits and cloud computing are explained.
3. Predictive Analytics is carried out on the public dataset, which helps to understand the failure patterns, and failure-related graphs are plotted.

This paper is structured as follows: Section 1 defines Big Data, the necessity of Big data analytics in manufacturing. Section 2 provides the literature review of various research papers published earlier. Section 3 provides a brief introduction to the Hadoop ecosystem which includes MapReduce, Spark, Hive, etc. Frameworks. Section 4 provides a brief introduction to cloud computing. Section 5 provides applications of Big data analytics in manufacturing along with its benefits. Section 6 provides our work in predictive maintenance. Finally, in Section 7 we explain the problem statement along with its results and various failure graphs.

A. Introduction to Big data:

The term 'Big Data' is not only defined by its Volume but there are several V's that define Big data. As mentioned in (4) we have 4 V's those are defined as follows:

1. **Volume**: It refers to vast amounts of data that is generated every second, minute, hour, and day in our digital world.

2. **Velocity**: It refers to the speed at which data is being generated and the place at which data moves from one point to another.
3. **Variety**: It refers to the ever-increasing different forms that data can come in such as text, images, videos, audios, and geospatial data.
4. **Value**: It refers to the importance of data for data analytics, value tells the motive behind the Data analytics.

Fig. 1.1 Big Data

There are additional V's along with above are mentioned in (3) are Veracity, Variability, Visualization.

5. **Veracity**: It refers to the quality and accuracy of the data. It highlights noise, biases, and abnormality in data.
6. **Variability**: It refers to data whose meaning is changing constantly.
7. **Visualization**: It refers to displaying data in charts, graphs, and any other visual form of data.

The data generated in three formats (2) from various sources are structured data, unstructured data, and semi-structured data; these are mainly generated from machines, social media, and organizations respectively.

B. Big Data analytics:

Data Analytics methods in manufacturing can be categorized into Descriptive Analytic, Diagnostic Analytics, Predictive Analytics, Perspective Analytics. They can describe as follows:

Descriptive Analytics:

An exploratory analysis is done on the previous collected data, with the intention of knowing what happened is 'descriptive analytics'. In this stage, we explore the data using data mining tools and statistics analysis methods like mean, mode, median, average, etc. used. Descriptive analytics focuses on what happened in past. The visualization is also used to understand the trends, patterns and the range of any particular or all variable in the dataset. this type of analytics can be used to manage and understand the life of manufacturing tool (5). It is used in capturing misbehavior patterns of product and tools.

Diagnostic analytics:

Diagnostic analytics and descriptive analytics both are done analysis of past collected data. Since, descriptive analytics finds out what happened, the diagnostic analytics finds out 'why happened', frequency of happenings can be found out in these analytics. In manufacturing tool case, we can find out that why tool fails in any particular operation (6). It is used in fault diagnostic and anomaly detections.

Predictive analytics:

Predictive analytics as name suggests it predicts what will happen in future. Analysis is done on previous data to see future trends in data. Predictive analytics help organization to predict the future sale of product either increase or decrease or predict the failure of any manufacturing tool in manufacturing industries. Predictive analytics uses visualization graphs, data mining, statistical modelling and machine learning to forecast future outcomes (2). It is used in maintenance of machines and customer behavior prediction.

Prescriptive analytics:

Perspective analytics is done to get to know what should be done to get expected outcomes. Results of all above analytics that is descriptive analytics, diagnostic analytics and predictive analytics are used in perspective analytics to get the result of perspective analytics. Perspective analytics uses all the tools used by the predictive analytics. (2). Perspective analytics is used in system reliability and optimization of system.

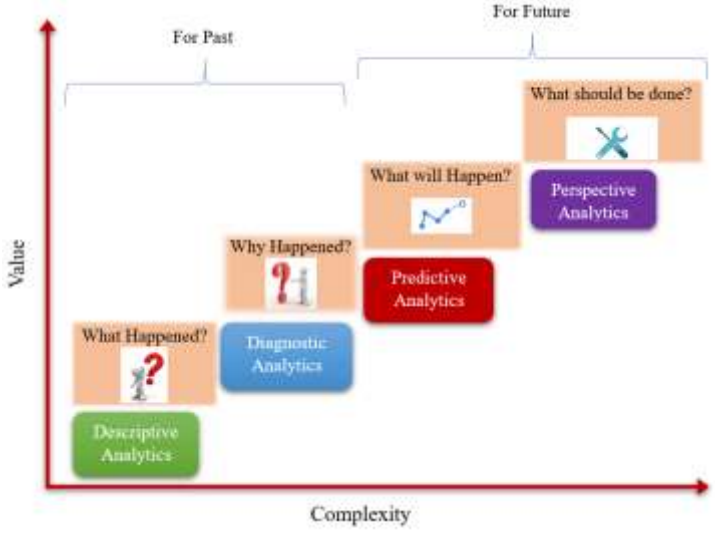


Fig.1.2 Types of Data Analytics (2)

C. The necessity of Big Data Analytics in Manufacturing:

The data to be analyzed are not just in large quantities, but they are composed of various data types. Since we get data from the machine, it's massive data, which has unique features, it comes in various formats, it might contain noisy samples or errors (7). Although we know that we can get valuable information or some knowledge from data, more data doesn't need to give more information. To analyze this huge data generated during various stages of manufacturing & to get some useful insights from it, it is first necessary to understand the necessity of Big Data analytics in manufacturing, as it may reveal some interesting patterns, which were not visible earlier

The necessity of Big Data analytics in manufacturing is as follows:

Improving factory operations and production: Product manufacturing data and customer demand data can be analyzed using predictive analytics which will help manufacturing enterprises to enhance the factory operations and to improve machine utilization (2). For example, the demand for a certain product like vehicles is often related to festivals or at the start of the new year, this can be referred to as thick data related to the emotions of people (8)**.** Forecasting these occasions will help to make early allocations of resources to meet the demands of customers.

Reducing Machine Downtime: In manufacturing enterprises, the prevalent sensors deployed throughout the whole production line on the shop floor can collect various data reflecting machinery status (2). For example, analysis of machine data would help us to understand the main cause of the failure of machine components, so we can reduce machine downtime. Earlier we used to use Fault Tree Analysis (FTA) charts to find out the fault and its root cause which was a very lengthy process and time-consuming (9).

Improving Product Quality: The customer needs or requirements and market demand could be useful to improve the product design for product improvement. To improve the quality of products, the quality department of manufacturing enterprises can share this data with the design team (10). During product manufacturing, we can reduce defective products by analyzing manufacturing data which will help to find out the root cause of the defect (11). As a result, the product quality can be improved.

Reducing Waste: Manufacturing industries produce sizable number of wastes like defecting product which includes scrap or rework**.** Along with defects manufacturing waste also includes overproduction, waiting, slow moving or obsolete inventory (SLOB), transportation. Earlier to reduce waste organizations have used 5-whys analysis and root cause problem solving (RCPS) methods (12). Since, organizations acquiring data from above mentioned sources, big data analytics can be used to understand the root cause of any waste problem through visualization graphs and machine learning model. This can reduce defecting products, improving their quality, reducing over production by forecasting customers demand (13).

Make or Buy Decision: Make or buy is a decision of an organizations about either manufacturing a particular product or service in the organization or buy it from outside suppliers. Earlier these decisions taken on the basis of quantitative analysis which considers cost of manufacturing (in company), material, labor, inventory, etc. and their benefits. Every time for quantitative analysis organizations have to spend lot of resources (Money and Time) (14). Nowadays companies are generating huge data which can be used to analyse the demand of any product, how can it be beneficial to organization if product is made in house or bought from external suppliers. (15). The analysis can be done using data visualization, statistical modeling and data mining. This will reduce the man power, money and time required to quantitative analysis.

Enhancing Supply Chain efficiency: The use of IoT devices like sensors, RFID tags in supplies, inventory stocks, during manufacturing, sales and dispatch of the final product, transportation of final product to customer produces massive data. All this data can be analyzed so we can reduce supply risk, find out optimal logistic routes, predict the delivery of the product to the customer, and so on (16). Furthermore, by analyzing inventory data, enterprises can reduce the cost of excessive holdings and can have optimistic safety stock in inventory (2). As a result, supply chain efficiency can be greatly improved.

Improving customer experience: Companies can obtain customer data from after-market sources, such as sales channels, partner distributors, retailers, social media platforms, direct feedback to the company, and review on E-commerce sites (16). Big data analytics can be used for carrying out a sentiment analysis on customer data (16) to understand the sentiment of people about a specific product, its design, and its reliability. With sentiment analysis, enterprises can improve customer experience about companies' products.

D. Maintenance strategies:

The main purpose of maintenance to either eliminate the failures or reduce them during production process. Maintenance can also be defined as, 'it is the process of restoring the machine to its initial state to fulfill the intended work'. The objective of maintenance is to select such maintenance process which will reduce or eliminate failure causes in cost-effective way (4). There are two types of maintenance, reactive (after failure) and proactive (before failure) maintenance. In proactive maintenance, preventive maintenance, condition-based maintenance and predictive maintenance are there, briefly explained:

Preventive Maintenance:

preventive maintenance is proactive maintenance and is performed after fixed interval of time. The primary purpose of preventive maintenance is to calibrate the machine to initial statues and correct any occurred or possible failure problems before breakdown of machine. Preventive maintenance includes inspection of machine and its components, basic repair work or replacement of any worn-out component, lubrication of friction affected parts, cleaning and adjustment of components. Basic preventive maintenance can be done by machine operator with proper training. Preventive maintenance is less costly compared to predictive maintenance (17).

Condition-Based Maintenance (CBM):

Condition based maintenance comes in between preventive maintenance and predictive maintenance (18) and (19). The primary aim of CBM is to monitor determined machine component and based on the monitored information recommend the maintenance decisions (18). The CBM performs real time monitoring of machine component to make right decision on right time by reducing unnecessary maintenance cost. Hence, replacement or repair of component done only when performance of that component is below than that of predetermined limits. These predetermined limits are decided based on previously collected and analyzed machine data. (18). CBM is preventive in approach and tells what will fail in near future. It needs humans to make decision about maintenance.

Predictive Maintenance:

Predictive maintenance predicts the failure of machine in future, so organizations can plan machine maintenance in advance. It also helps to reduce frequent unplanned downtimes with better managing inventory and maximizing equipment lifetime. Predictive maintenance provides exact time for scheduled maintenance (4). The basic flow of predictive maintenance is shown in below figure:

Fig.1.3 predictive maintenance workflow

In predictive maintenance, first we acquire machine data using IoT devices like sensors, gauges, thermometers, etc. this machine data will have healthy as well as faulty data. In the case of manufacturing machine data, it will have operational data like speed, feed, depth of cut, etc., material data like hardness, toughness, wear resistance, etc., (20). After that this machine data preprocessed to remove abnormal values and outliers.

In condition indicator, all the influencing variables are extracted and selected for further process. These selected independent variables are called features and independent variables are called label. Based these data a perfect fit machine learning model is trained and deployed in predictive maintenance.

II. LITERATURE REVIEW:

In this section, the past research efforts on Big data analytics presented, what are the methods for analysis, and the evaluation metrics used to evaluate the prediction algorithms. Previous work on Big data analytics:

The data can be analyzed in four ways, all these types are discussed in (2). Along with the analysis method, this paper also discusses what are the necessity and challenges of Big Data Analytics in manufacturing. Then, the enabling technologies of big data analytics in manufacturing are also discussed.

The adoption and use of 'Internet of Things (IoT) technology is leading us to the fourth industrial revolution. Paper (1) discusses the idea of using IoT in the industrial area. This paper introduces the Industrial Internet of Things (IIoT). This technology will lead us to Industrial Automation and control systems (IACS). This paper defines IIoT, its use in the manufacturing field, which can be used in real-time analysis of machine failure.

As most of the industries use the robotic arm for manufacturing products in industry and IoT which facilitates machine to machine communication, the paper (16) discusses an idea of making and using the Internet of Robotic Things (IoRT). This will help analyzers to obtain accurate and continuous data from the machine in real-time or batches, which can be used for analysis for predictive maintenance with the help of data analytics.

Since the age of Big Data is coming, the traditional data analysis method may not work as it used to, since the size of the data is too large. (7) discusses how to develop a platform to analyze big data. This paper also introduces us to Data Analytics and Big Data Analytics. This paper discusses the process of data mining and names this process as the process of knowledge discovery. It also gives a brief introduction to the evaluation of the process.

Big Data analytics is a challenging and time-consuming task, and it requires a large computational infrastructure. (21) discusses various big data frameworks like Hadoop and Spark, along with this it also discusses the methods of data pre-processing for smooth analysis of data. There are many challenges while processing the data, all these challenges and new possibilities are discussed in this paper.

Predictive analytics is a type of Big Data analytics as mentioned in the (4) paper. This paper explains how we can use big data analytics for predictive maintenance, this paper also discusses other maintenance strategies and how we can use big data for predictive maintenance. This paper also explains big data architecture and big data analytics framework for predictive maintenance.

Paper (1) talks about how we can use IoT technology in the manufacturing field. But there are some challenges for implementing IoT in the industrial area, all these challenges are discussed in (22) survey. This survey also discusses five important methodologies of industrial big data analytics. It also talks about applications of industrial big data. To analyze big data, we use cloud computing and edge computing methods. (23) this article highlights the core applications of edge computing. It also discusses the importance of edge computing for Industrial big data analytics.

Today is the era of cloud computing technology where big data technology is used. The authors of (24) review papers claim they have reviewed more than 30 articles to understand what cloud computing is and to evaluate the meaning and importance of cloud computing. They give their conclusion before cloud computing and after.

As the above paper defines what is cloud computing, (25) in this article the main characteristics of cloud computing are discussed. Benefits and limitations are also discussed. The authors of the research paper aim to assist academic researchers and manufacturing enterprises.

The paper presented by (26) mostly talks about the bottleneck production system. In this paper, the author mentions the ways to evaluate ML algorithms. The confusion matrix, its definition, how we can calculate the confusion matrix all are discussed in this paper.

Paper (3), also talks about evaluation, but this discusses more on the characteristics of Big Data and Big data analytics. This research paper gives brief info about the platforms that can be used for big data analytics and the issues with the existing system.

III. HADOOP ECOSYSTEM OVERVIEW:

Manufacturing Enterprises produce a huge amount of data i.e., Big Data. To process this data, the Hadoop platform is available. Apache Hadoop is an open-source framework that makes it easy to interact with Big Data. Hadoop can store and process a vast amount of data using commodity cluster hardware. This capability of HADOOP can be leveraged to analyze voluminous amounts & a variety of data generated during various stages of manufacturing. Hadoop has two main components that are specifically designed to work with big data. Those are 1. Hadoop Distributed File System (HDFS), stores data in a distributed way. 2. MapReduce processes data in a parallel way in clusters. The following figure shows the Hadoop ecosystem:

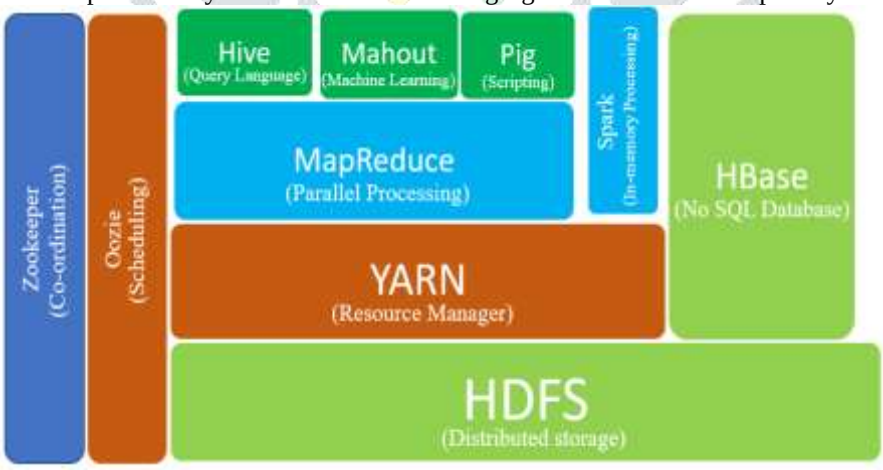


Fig.3.1 Hadoop Ecosystem

As shown in the figure, there are more components in the figure other than HDFS and MapReduce, we call these components like Hive, Pig, Oozie, Zookeeper, HBase, Mahout as Hadoop projects (27). These projects help Hadoop to write easy codes and speed up the process of processing in a cost-effective manner. All the Hadoop and its components are discussed below:

Hadoop distributed file system (HDFS): HDFS is a storage system for the Hadoop ecosystem. As the name suggests, it stores data files in a distributed way. HDFS is mainly designed to store a vast amount of data usually in petabytes or more (28).

MapReduce: MapReduce is one of the most important components of Hadoop. MapReduce does use distributed and Parallel algorithms for processing big data. MapReduce makes use of two functions: Map and Reduce (29) Both the operation map and reduce are performed on the data nodes. MapReduce takes input from HDFS and stores output back to the HDFS (29).

YARN: YARN is also called Hadoop version 2. YARN is basically a resource (Ram and CPU) manager for Hadoop. It interacts with applications and schedules resources for their use (28) during processing.

SPARK: Apache Spark is often considered as Hadoop version three. It has no own storage system; it uses HDFS to store its data. Spark can also use YARN as a resource manager. It is a platform that handles all the process consumptive tasks: batch processing, interactive or iterative real-time processing, graph conversion, and visualization, etc. (28). It consumes in-memory resources hence; it is faster than the prior in terms of optimization. Spark is the best suited for real-time data whereas Hadoop is the best suited for structured data or batch processing (30).

Hive: Apache hive is an open-source data warehouse software for reading, writing, and managing large data set files that are stored directly in HDFS or HBase (27). The data warehouse system used to summarize, analyze, and query the data of larger amounts in the Hadoop platform is called Hive.

PIG: Pig is a scripting language built on the top of MapReduce. It was developed by the YAHOO which works on a pig Latin language, which is a query-based language similar to SQL (27).

MAHOUT: Mahout is built on the top of MapReduce and it is the best tool for Machine learning. It has a java machine learning library. It does include algorithms like clustering, classification, and collaborative filtering (31).

HBASE: HBase is a NoSQL database, built on top of the HDFS. HBase is scalable and uses distributed ways to store data. It supports all types of data like structured, unstructured, and semi-structured data (27).

OOZIE: Oozie acts as a scheduler for the Hadoop ecosystem. Its simple task is to schedule the job to each component like Hive, Pig, HBase, MapReduce, etc. oozie facilitates scheduling of jobs that need to run on a fixed time.

ZOOKEEPER: Zookeeper is a centralized management system for synchronization. It gives greater coordination between Hadoop components and ensures the High availability of running tools (32).

IV. APPLICATIONS AND BENEFITS OF BIG DATA ANALYTICS IN MANUFACTURING:

Fig.4.1 Big Data Analytics Applications in Manufacturing

Big data analytics can be used in predictive maintenance (4), predictive quality (33), product lifecycle management (34) fault detection (11), tool life cycle optimization (35), supply chain management (16) , sentiment analysis (36) , etc. Big data analytics offers various benefits for given applications, as follows:

1. Minimizes machine downtime and maintenance cost (37).
2. Early fault detection prevents abnormal machine failures (11).
3. Optimizes the process of the supply chain (16).
4. Gain the competitive edge in the shortest possible time (36).

V. CLOUD COMPUTING:

Cloud computing is a technology that allows users to use the servers provided by the special service providers via the internet to store and process the data instead of using local computing machines (38). There is a cloud service provider, they solely provide this service to other business users. These users have to pay for the service they use. The cloud services offer smooth running of a local computer system by handling all of their workloads. There are three services provided by cloud computing 1. Software as a service (SaaS) 2. Platform as a service (PaaS) 3. Infrastructure as a service (IaaS) (24).

Cloud computing platforms available for Big Data Analytics: amazon web services (AWS), Microsoft Azure, IBM Cloud, Google Cloud, Alibaba Cloud, Oracle, Databricks, etc. out of these services we have used Databricks environment for conducting this research work. These cloud computing offers benefits like agility, scalability, flexibility, reduces the working load from the local system (25).

VI. METHODOLOGY:

Refer Fig 6.1 flowchart shows how the workflow goes while analyzing the data.

As we are analyzing big data in manufacturing, we have used a cloud computing provided by Databricks environment to process the data. This environment allows code using python and it provides a framework called PySpark, where we can use python as well as SQL to explore the data, however, the only python can be used to code ML programs. So, the first step of data analysis is to acquire data from machines and store it in big data storage. In this case we have stored data on Databricks storage.

Firstly, the data from distributed storage is imported to the processing unit. Subsequently, importing the most important stage is 'data exploration'. In data exploration, about 30% of the time is spent understanding the data, understanding what kind of values are there. We need to know the number of unique values. In this stage, we find out the null/Nan values. We also find out the outliers. Since I have used the Apache Spark framework to write my code, it allows me to write python code. For data

exploration, we have used both the languages first python and then SQL which means we can use both languages only to explore data for understanding purposes.

After exploring the data, we understood there are null values and unique values other than numbers. The values other than numbers need to be removed from the data. We have many variables, having more variables does not mean that all the variables are useful for us. Some of them are just to give more information to us, we cannot use such a type of variable for the ML model. In our data, we have UDI which shows the serial number and product ID is a unique ID given to each product. If we use such a variable while processing it may affect the accuracy of model. Now the next stage is one of the most important. In this stage, our task is to find out the dependent and independent variables from the data. We name them features and labels. A label is a target column or simply an independent variable and features are the dependent variable. While exploring the data, it becomes necessary to figure out the relation between features and label columns. So, we can find out whether any feature makes maximum impact on the label. We can do it by visualization method using graphs.

We have selected features based on failure of Machine component. Variables like 'tool wear' and 'rotational speed' are directly responsible for 'tool wear failure' and 'heat dissipation failure' respectively. However, there are other hidden features responsible for failure of machine components. Therefore, we have extracted features like 'temperature difference', 'power' product of torque (Nm) and rotational speed (rad/sec), 'tool wear torque' which is product of tool wear (min) and toque (Nm). All the failures, occurred due to the readily available and extracted features are explained in problem statement section with Box-graphs and number of failures occurred along with their failure range.

We find out features label columns and their relationships with each other and split the data into train and test sets. Splitting of data can be done in 70-30 or 80-20 criteria where 70 or 80 % data is considered as a training set and the rest is for evaluation for machine learning. In this case, 70-30 is considered.

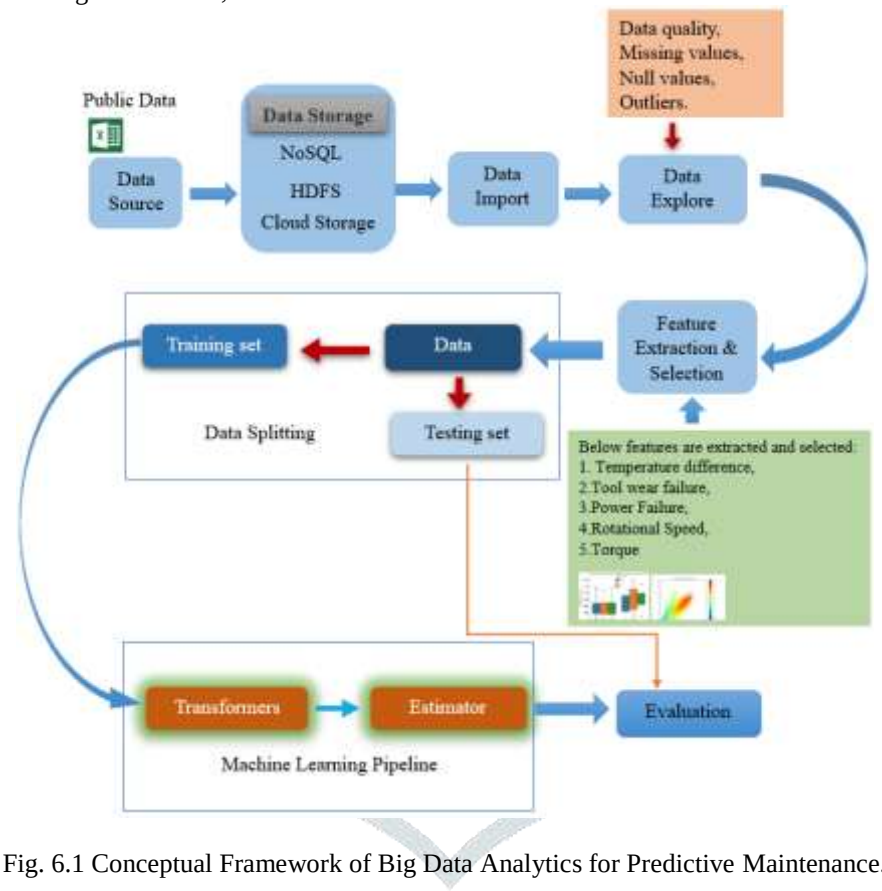


Fig. 6.1 Conceptual Framework of Big Data Analytics for Predictive Maintenance.

As the dataset has some missing values, it is necessary to either remove them or fill them with the proper value. If the missing values are not in large numbers, then we can remove them. But, in this case, we have a large number of missing values nearly 2000 in entire dataset, removal of these values will have an impact on model accuracy, hence we have to fill it with proper values like mean or median. The process of filling the missing value is called the 'Imputation process'. We have two data types here in the dataset, one is float and the other is a category. So, first, we will find out the categorical and numerical variables. Subsequently, finding the data types variables column-wise, we create an ML Pipeline.

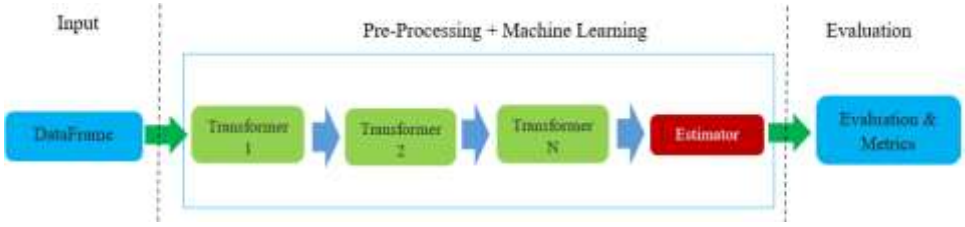


Fig.6.2 Spark Machine learning pipeline

In this Pipeline, we can have different stages called one or more than one transformer. These transformers are mainly used for data pre-processing. In the data, there are null values. Therefore,
Transformer 1: this transformer imputes missing and null values in data.

Transformer 2: the second transformer is one hot encoder; it converts categorical values into zeros and one.

These two main transformers are used during analysis here in this project. The Transformer N represents that, we can use N number of transformers in a machine learning pipeline depending upon the user's requirement.

Input of the transformer is DataFrame and its output is also a DataFrame. In the spark ML pipeline, after creating a Number of transformers, we have to create an estimator. In the estimator stage we create an ML model, we have selected classification models, logistic regression to be specific. An estimator trains a model using training data. First, we fit the data into the model, then we transform data. Transforming data is understanding the data, how features and label columns are related to each other.

In evaluation, which is the last stage of the flowchart, we evaluate our model by calculating its accuracy, precision, confusion matrix, etc. using a testing dataset. Evaluation gives insight into the model accuracy, which tells the reliability of the model. We can decide if it is an overfitted model or under the fitted model or perfect fit model.

Once we are done with model creation, we can deploy that production model. To deploy our ML model, we must have one web development application and a server. To deploy the model on the server we have to purchase a server. However, we can deploy the model on the local server. We can use 'Flask' or 'Django' as a web development application.

VII. PROBLEM STATEMENT:

We have taken a public dataset (39) with 10,000 samples and it has 14 variables.

Table 7.1 Dataset:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	UDI	Product ID	Type	Air temperature [K]	Process temperature [K]	Rotational speed [rpm]	Torque [Nm]	Tool wear [min]	Machine failure	TWF	HDF	PWF	OSF	RNF
2	1	M14860	M	298.1	308.6	1551	42.8	0	0	0	0	0	0	0
3	2	L47181	L	298.2	308.7	1408	46.3	3	0	0	0	0	0	0
4	3	L47182	L	298.1	308.5	1498	49.4	5	0	0	0	0	0	0
5	4	L47183	L	298.2	308.6	1433	39.5	7	0	0	0	0	0	0
6	5	L47184	L	298.2	308.7	1408	40	9	0	0	0	0	0	0
7	6	M14865	M	298.1	308.6	1425	41.9	11	0	0	0	0	0	0
8	7	L47186	L	298.1	308.6	1558	42.4	14	0	0	0	0	0	0
9	8	L47187	L	298.1	308.6	1527	40.2	16	0	0	0	0	0	0
10	9	M14868	M	298.3	308.7	1667	28.6	18	0	0	0	0	0	0

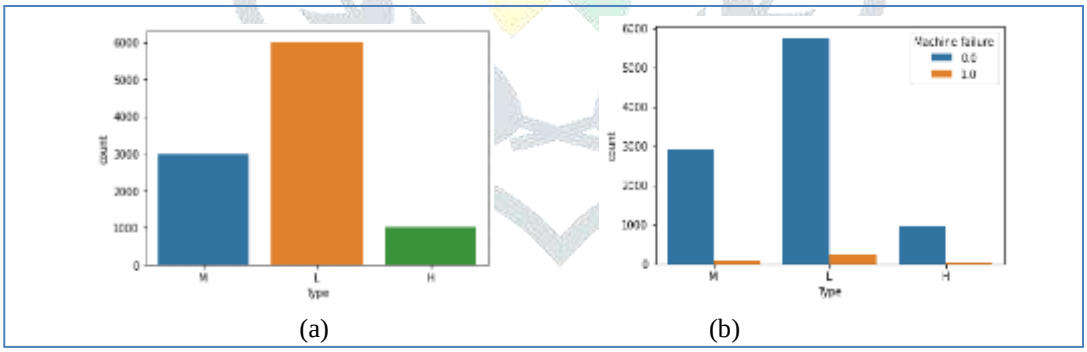
In this table 7.1 data set, we can figure out what is the meaning of data for most of the columns.
Type = Type is basically showing the quality of product, L = low, M = medium, H = High.
TWF = Tool wear failure, HDF = Heat Dissipation Failure, PWF = Power Failure, OSF = Overstrain Failure, RNF = Random Failure

For columns from Machine failure to RNF, left to right, we have samples in the form of 0 and 1. This '0' and '1' indicates No failure and failure of the machine respectively. The machine fails whenever any of the TWF, HDF, PWF, ODF, RNF fails.

A. Data Exploration:

1.**Product and machine failure count by their quality:**

From Below figure 7.1, (a) one can understand that number of products by their product quality. There are 999 products in high quality, 5995 in low quality, 2986 in medium quality. Fig.7.1 (b), shows the quality wise failures in product manufacturing. This will make supervisor or maintenance department learn there are more failures in the low-quality product about 235 failures than rest together (H=21, M=83). Hence, they can ensure less failure during low quality product failure.



(a) (b)

Fig.7.1 counting by (a) machine failure (b) type of quality.

2.**Failure by Tool Wear Failure:**

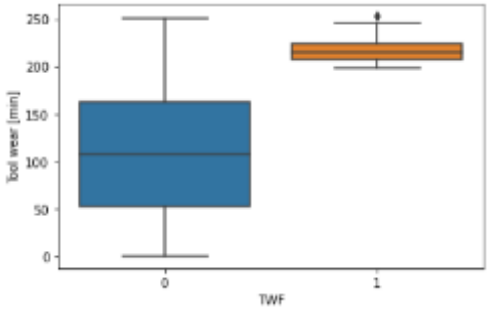
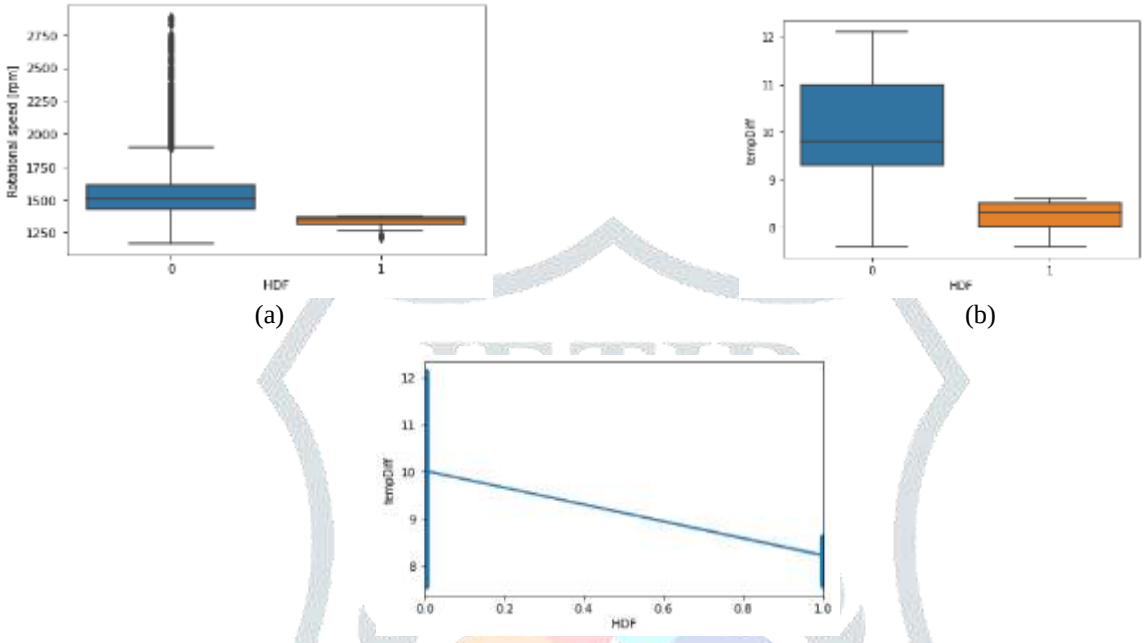


Fig. 7.2 Graph for Tool Wear Failure.

From figure 7.2, machine operator who is responsible for operations on machine can understand that if tool wear is more than 200 minutes it will cause machine failure. The machine has failed 46 times. From these machine operators will understand that use of tool should be less than 200 minutes, so failure can be avoided or reduced. This will help to reduce unexpected machine downtime and improve quality of product.

3.**Heat dissipation failure:**

In heat dissipation failure occurs because of two variables namely rotational speed and temperature difference which leads to machine failure. The below graphs will make maintenance department and operator understand the failure of machine. From fig 7.3(a), first when the rotational speed is below nearly 1400 rpm, and from fig 7.3(b)&(c) second when the temperature difference between air temperature and process temperature is below 9 K (or nearly 8.5 K from regression graph(c)). There are 115 heat dissipation failure in the machine. Once operator understand the working range of the machine, so they can ensure rotation speed and temperature difference in the working limits.



(a) (b)

(c)

Fig. 7.3 Graphs for heat dissipation failure.

4.**Power failure:**

Machine has failed 95 time in total citing power failure reason. From figure 7.4, machine operator and supervisor can understand that the working range of power is approximately in between 35,000 watt and 90,000watt. If power value is not in limits, it will make machine fail. To avoid power failure operator should keep machine power in between range.

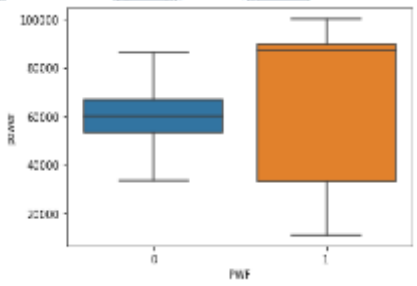


Fig. 7.4 Graph for Power Failure

5.**Over strain failure:**

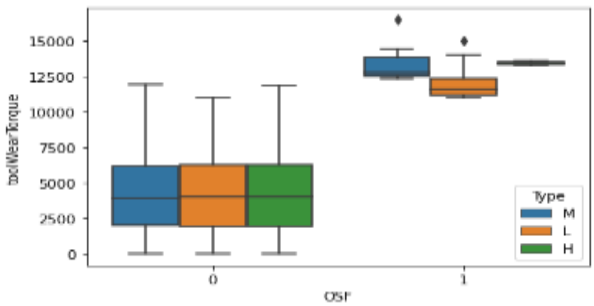


Fig. 7.5 overstain failure

Over strain failure occurs when the product of tool wear and torque goes beyond limit. From Figure 7.5 operator can understand that; the over strain failure does not occur uniformly on every product but it is depending on Type of Quality. The working range

for all types is approximately same. The machine has failed 98 times because of overstrain failure. For low quality product when the product of tool wear and torque goes beyond limit of approximately 11,000 Nm-min, for medium quality approximately 12,500 Nm-min and for high quality approximately 13,500 Nm-min. operator can keep product of tool wear and torque in specified limit for each quality of product to avoid overstain failure.

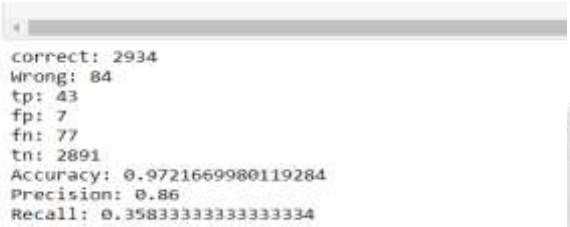
6.**Random Failure**: The random failure is not related to any kind of parameter. This failure occurs randomly and it is not depending on variable.

B. Accuracy of the ML models:

When we create any ML model, it is necessary to find out that the model we have is accurate enough so we can deploy it in production using any web application. To evaluate the model, we will use a confusion matrix. We have created two machine learning models and will compare them with each other to see which is a more suitable model.

1. **Logistic Regression model:**
   Confusion matrix:

2. **Random Forest model:**
   Confusion matrix:



VIII. CONCLUSION:

In this paper, we have given an insight into how enterprises can use big data analysis in manufacturing. Moreover, we have explained big data, its sources, and format, the necessity of Big data analytics in manufacturing.

- There are four types of data analytics namely descriptive, diagnostic, predictive and perspective analytics all explained briefly. And also, maintenance strategy like preventive maintenance, condition-based maintenance and predictive maintenance are discussed.
- Also, the Hadoop ecosystem briefly explained. The plethora of available platforms for big data analytics over cloud are explained in this paper. So, the manufacturing enterprises will have opportunity to use this new breed of the software for handling the complex number of datasets which was not possible with traditional approach.
- Further, the Big data analytics leveraged for predictive maintenance. The problems manufacturing enterprises face in maintenance can be solved using Big data analytics and can have better utilization of resources, reduction in unplanned downtimes, improvement in product quality.
- Manufacturing enterprises can see the analytical visibility about the predictive analytics and many other events on the dashboards which was not possible earlier. And hence, enterprises can take precautionary measures and avoid unplanned downtimes.
- Likewise, Big data analytics can be leveraged for sentiment analysis in sales and marketing to understand the emotions of customers towards their products. Big data analytics can also be leveraged in Enhancing supply chain management.

Since, machine either fails or doesn't fail which is converted to the binary '0' and '1', considering this we have selected model Logistic regression and Random forest classifier belongs to binary classification category. To compare these two model we haves used accuracy, precision and confusion matrix. From the results, the accuracy of the logistic model is less than that of random forest classifier however precision of logistic model is more.

IX. REFERENCES

1. **Boyes, Hugh , et al.** *The industrial internet of things (IIoT): An analysis framework.* Coventry : Elsevier B.V., 2018.
2. **Dai, Hong-Ning, Wang, Hao and Xu, Guangquan.** *Big Data Analytics for Manufacturing Internet of Things: Opportunities, Challenges and Enabling Technologies.* 2019.
3. **Khan, Mudassir.** *Big Data Analytics Evaluation.* Abha : ResearchGate, 2018.
4. **C. K. M., Lee, and Yi Cao, Kam Hung Ng.** *Big Data Analytics for Predictive Maintenance Strategies.* s.l. : Research Gate, 2017.

5. **UNSW, Sudney.** Descriptive, Predictive & Prescriptive Analytics: What are the differences? *The university of new south wales website.* [Online] The university of new south wales , January 29, 2020. https://studyonline.unsw.edu.au/blog/descriptive-predictive-prescriptive-analytics.
6. **Geeks, Geeks for.** Types of Analytics . *Geeks for Geeks.* [Online] Geeks for Geeks, January 19, 2021. https://www.geeksforgeeks.org/types-of-analytics/#:~:text=Descriptive%20Analytics%20%3A,it%20predicts%20the%20future%20outcome..
7. **Chun-Wei, Tsai, et al.** *Big data analytics: a survey.* s.l. : Springer, 2016.
8. **MacMillan, Andy.** Why Big data isn't enough for businesses - They need thick data too. *The Economics Times.* 2020.
9. **S.S., Rao and Singiresu, S.** *Engineering Optimization: Theory and Practice.* s.l. : John Wiley and sons, 2009.
10. **Ji-hye, Jun, Tai-Woo, Chang and Sungbum, Jun.** *Quality Prediction and Yield Improvement in Process Manufacturing Based on Data Analytics.* s.l. : MDPI, 2020.
11. **Ndeye Gueye, Lo, Jean-Marie, Flaus and Olivier, Adrot.** *Review of Machine Learning Approaches In Fault Diagnosis applied to IoT System.* s.l. : HAL, 2019.
12. *Scrap loss reduction using 5-whys analysis.* **Benjamin, Samuel Jebaraj, Marathamathu, Srikamaladeci M. and Muthaiyah, Saravanan.** 5, International Journal of Quality and Reliability Management : s.n., 2009, Vol. 27.
13. *Data Analytics Platform for the Optimization Waste Management Procedures.* **Thanasis, Vafeiadis, et al.** s.l. : Research Gate, 2020.
14. **Thakur, Madhuri and Vidya, Dheeraj.** Make or Buy Decision. *WallStreetMojo.* [Online] WallStreetMojo. https://www.wallstreetmojo.com/make-or-buy-decision/.
15. *The Importance of Big Data Analytics in Business: A case study.* **Hiba, Alsghaier, et al.** 4, s.l. : American Journal of Software Engineering and Applications., 2017, Vol. 6.
16. **Mahya, Seyedan and Fereshteh, Mafakheri.** *predictive big data analytics for supply chain demand forecasting: methods, applications and research opportunities.* s.l. : Springer, 2020.
17. *Preventve Maintenance (PM) Planning: Review paper.* **Bansari, Emnie illyani and Razak, Hamid abdul.** 2, s.l. : Journal of Quality in Maintenance Engineering, 2017, Vol. 23.
18. *Condition-based maintenance implementation: a literature review.* **Teixeira, Humberto Numo, Braga , Cristina and Lopes, Isabel.** Athens : Science Direct, 2021. 30th International conference on flexible automation and intelligent manufacturing (FAIM2021).
19. *A review of condition-based maintenance decision-making.* **Ahmad, Rosmaini and Kamaruddin, Shahrul.** s.l. : Research Gate, 2012. European J. Induatrial Engineering . Vol. 6.
20. *Machine learning Framework for predicitve maintenance in milling.* **Traini, Emiliano, et al.** s.l. : Science direct, 2019. International Federation of Automatic Control .
21. **Salvador, García, et al.** *Big data preprocessing: methods and prospects.* s.l. : Creative commons, 2016.
22. **JunPing, Wang, et al.** *Industrial Big Data Analytics: Challenges, Methodologies, and Applications.* s.l. : Research Gate, 2018.
23. **khan, Wazir Zada, et al.** *Edge computing: A survey.* s.l. : Research Gate, 2019.
24. **srivastava, Priyanshu and Khan, Rizwan.** *A Review Paper on Cloud Computing.* s.l. : Research Gate, 2018.
25. **Wang, Peng, Gao, Robert X. and fan, Zhaoyan.** *Cloud Computing for Cloud Manufacturing: Benefits and Limitations.* s.l. : Research Gate, 2015.
26. **Subramaniyan , Mukund, et al.** *A data-driven algorithm to predict throughput bottlenecks in a production system based on active periods of the machines.* s.l. : Elsevier Ltd, 2018.
27. **uzunkaya, can, Ensari, Tolga and Kavurucu, Yusuf.** *Hadoop Ecosystem and its Analysis on Tweets.* s.l. : Science DIrect.
28. **landset, Sara, et al.** *A survey of open source tools for machine learning with big data in the hadoop ecosystem.* s.l. : Springer, 2015.
29. **Nagdive, Ashlesha S. and Tungnayat, R.M.** *A Review on Hadoop Ecosystem for Big Data.* s.l. : Research Gate, 2018.
30. **shaikh, Eman, et al.** *Apachr Spark: A Big Data Processing Engine.* s.l. : second IEEE Middle East and North Africa communication conference, 2019.
31. **Seminario, Carlos E. and Wilson, David C.** *Case study Evaluation of Mahout as a recommender Platform.* s.l. : Research Gate, 2012.
32. *Big Data and Business Analytics: Trends, Platforms, Success Factores and Applications.* **Ajah, Ifeyinwa Angela and Nweke, Henry Friday.** 32, s.l. : Big data and cognitive computing , 2019, Vol. 3.
33. **Miljikovic, Dubravko.** *Fault detection methods: A literature survey.* s.l. : Research Gate, 2011.
34. **wynn, Martin george.** *product life cycle management systems and business process improvement : A report on case study research.* s.l. : Research Gate, 2008.
35. **Karandikar, Jaydeep.** *Machine learning classification for tool modeling using production shop floor tool wear data.* s.l. : Science Direct, 2019.
36. **Shathik, Anvar and Prasad, Krishna.** *A Literature Review on Application of Sentiment Analysis Using Machine Learning Techniques.* s.l. : Research Gate, 2020.
37. **Paolanti, Marina, Felicetti, Andrea and Mancini, Adriano.** *Machine Learning Approach for Predictive Maintenance in Industry 4.0.* s.l. : Research Gate, 2018.
38. *Cloud Computing Basics.* **SRINIVAS, J., SUBBA REDDY, K.VENKATA and QYSER, Dr.A.MOIZ.** 5, 2012, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 1, p. 6.
39. **california, University of.** *Machine Learning Repository* . [Online] University of california . https://archive.ics.uci.edu/ml/datasets/AI4I+2020+Predictive+Maintenance+Dataset .